

# A General Framework for Transmission with Transceiver Distortion and Some Applications

Wenyi Zhang, *Member, IEEE*

## Abstract

A general theoretical framework is presented for analyzing information transmission over Gaussian channels with memoryless transceiver distortion, which encompasses various nonlinear distortion models including transmit-side clipping, receive-side analog-to-digital conversion, and others. The framework is based on the so-called generalized mutual information (GMI), and the analysis in particular benefits from the setup of Gaussian codebook ensemble and nearest-neighbor decoding, for which it is established that the GMI takes a general form analogous to the channel capacity of undistorted Gaussian channels, with a reduced “effective” signal-to-noise ratio (SNR) that depends on the nominal SNR and the distortion model. When applied to specific distortion models, an array of results of engineering relevance is obtained. For channels with transmit-side distortion only, it is shown that a conventional approach, which treats the distorted signal as the sum of the original signal part and a uncorrelated distortion part, achieves the GMI. For channels with output quantization, closed-form expressions are obtained for the effective SNR and the GMI, and related optimization problems are formulated and solved for quantizer design. Finally, super-Nyquist sampling is analyzed within the general framework, and it is shown that sampling beyond the Nyquist rate increases the GMI for all SNR. For example, with a binary symmetric output quantization, information rates exceeding one bit per channel use are achievable by sampling the output at four times the Nyquist rate.

## Index Terms

Analog-to-digital conversion, generalized mutual information, nearest-neighbor decoding, quantization, super-Nyquist sampling, transceiver distortion

W. Zhang is with Department of Electronic Engineering and Information Science, University of Science and Technology of China, Hefei, China. (e-mail: wenyizha@ustc.edu.cn)

This research has been funded in part by National Science Foundation of China through grant 61071095, Research Fund for the Doctoral Program of Higher Education of China through grant 20103402120023, and the Fundamental Research Funds for the Central Universities of China.

## I. INTRODUCTION

In digital communication systems, various forms of distortion are ubiquitous, acting as the main limiting factor for information transmission. Those distortions that come with the propagation of signal, such as shadowing and multipath fading, have received extensive research since the earliest era of digital communications [1]. The current paper, alternatively, concerns with the other category of distortions that come mainly with the engineering of transceivers. This category of distortions encompasses a number of models of practical importance, including the clipping or saturation of transmitted waveforms due to power amplifier nonlinearity, the analog-to-digital conversion (*i.e.*, quantization) of received samples, and others. Such distortions are difficult to eliminate, and indeed people may deliberately introduce them, for practical reasons like hardware cost reduction and energy efficiency improvement.

We can usually approximate the aforementioned transceiver distortions as memoryless deterministic functions. Those functions, however, are generally nonlinear operations and thus break down the linearity in Gaussian channels. From a pure information-theoretic perspective, nonlinearity may not impose fundamental difficulty to our conceptual understanding, since the channel capacity is still the maximum of mutual information between the channel input and the distorted channel output. From an engineering perspective, however, the general mutual information maximization problem is usually less satisfactory in generating insights, especially when such maximization problems are analytically difficult, or even intractable, for general nonlinear channel models.

There are a number of existing works that seek to characterize the information-theoretic behavior of nonlinear transceiver distortion, largely scattered in the literature. In [2], the authors examined the channel capacity of clipped orthogonal frequency-division multiplexing (OFDM) systems, with the key approximation that the distortion due to clipping acts as an additional Gaussian noise. Such an approximation originates from a theorem due to Busgang [3], which implies that the output process of a Gaussian input process through a memoryless distortion device is the sum of a scaled input process and a distortion process which is uncorrelated with the input process. Regarding Nyquist-sampled real Gaussian channels with output quantization, an earlier study [4] examined the achievable mutual information as the signal-to-noise ratio (SNR) decreases toward zero. Specifically, the numerical study therein revealed that for a binary

symmetric output quantizer, the ratio between the capacity per channel use (c.u.) and the SNR approaches  $1/\pi$ , and that for a uniform octal (*i.e.*, 8-level) output quantizer, this ratio is no less than 0.475. In [5], the authors further established some general results for Nyquist-sampled real Gaussian channels, asserting that with a  $K$ -level output quantization, the capacity is achieved by choosing no more than  $(K+1)$  input levels, and that with a binary symmetric output quantization the capacity is indeed achieved by using a binary symmetric input distribution. For  $K > 2$ , however, it is necessary to use intensive numerical methods like the cutting-plane algorithm to compute the capacity. The authors of [6] addressed the capacity of multiple-input-multiple-output block-Rayleigh fading channels with binary symmetric output quantization. In [7], the authors went beyond the Nyquist-sampled channel model, demonstrating that the low-SNR capacity of a real Gaussian channel with binary symmetric output quantization, when sampled at twice the Nyquist rate, is higher than that when sampled at the Nyquist rate. In [8], the authors proved that by using a binary asymmetric output quantizer design, it is possible to achieve the low-SNR asymptotic capacity without output quantization.

Recognizing the challenge in working with channel capacity directly, we take an alternative route that seeks to characterize achievable information rates for certain specific encoding/decoding scheme. As the starting point of our study, in the current paper we consider a real Gaussian channel with general transceiver distortion, and focus on the Gaussian codebook ensemble and the nearest-neighbor decoding rule. We use the so-called generalized mutual information (GMI) [9], [10] to characterize the achievable information rate. As a performance measure for mismatched decoding, GMI has proved convenient and useful in several other scenarios such as multipath fading channels [10]. Herein, in our exercise with GMI, we aim at providing key engineering insights into the understanding and design of transceivers with nonlinearity. The nature of our approach is somewhat similar to that of [11], where the authors addressed the decoder design with a finite resolution constraint, using a performance metric akin to cutoff rate that also derives from a random-coding argument.

The motivation for using the performance measure of GMI and the Gaussian codebook ensemble coupled with the nearest-neighbor decoding is two-fold. On one hand, such an approach enables us to obtain an array of analytical results that are both convenient and insightful, and bears an “operational” meaning in that the resulting GMI is achievable, by the specific encoding/decoding scheme whose implementation does not heavily depend on the nonlinear

distortion model. On the other hand, Gaussian codebook ensemble is a reasonable model for approximating the transmitted signals in many modern communication systems, in particular, those that employ higher-order modulation or multicarrier techniques like OFDM<sup>1</sup>; and the nearest-neighbor decoding rule is also a frequently encountered solution in practice which is usually easier to implement than maximum-likelihood decoding, for channels with nonlinear characteristics. Nevertheless, we need to keep in mind that compared with capacity, the performance loss of GMI due to the inherently suboptimal encoding/decoding scheme used may not be negligible.

The central result in the current paper is a GMI formula, taking the form of  $(1/2) \log(1 + \text{SNR}_e)$ , for real Gaussian channels<sup>2</sup> with general transceiver distortion. Here  $\text{SNR}_e$  depends on the nominal SNR and the transceiver nonlinearity, and we may interpret it as the “effective SNR”, due to its apparent similarity with the role of SNR in the capacity formula for Gaussian channels without distortion. The parameter  $\text{SNR}_e$  thus serves as a single-valued performance indicator, based on which we can, in a unified fashion, analyze the behavior of given transceivers, compare different distortion models, and optimize transceiver design.

Applying the aforementioned general GMI formula to specific distortion models, we obtain an array of results that are of engineering relevance. First, when the nonlinear distortion occurs at the transmitter only, we show that the Bussgang decomposition, which represents a received signal as the sum of a scaled input signal part and a distortion part which is uncorrelated with the input signal, is consistent with the GMI-maximizing nearest-neighbor decoding rule. This result validates the Gaussian clipping noise approximation for transmit-side clipping, as followed by the authors of [2].

Second, we evaluate the GMI for Nyquist-sampled channels with output quantization. For binary symmetric quantization, we find that the low-SNR asymptotic GMI coincides with the channel capacity. This observation is somewhat surprising, since the GMI is with respect to a suboptimal input distribution, namely the Gaussian codebook ensemble. On the other hand, there exists a gap between high-SNR asymptotic GMI and the channel capacity, revealing the penalty

<sup>1</sup>In the current paper we confine ourselves to the single-carrier real Gaussian channel model, and will treat multicarrier transmission with nonlinear distortion in a separate work.

<sup>2</sup>For complex Gaussian channels we also have an analogous result; see Supplementary Material VII-C.

of suboptimal input distribution when the effect of noise is negligible. For symmetric quantizers with more than two quantization levels, we formulate a quantizer optimization problem that yields the maximum GMI, and present numerical results for uniform and optimized quantizers. As an example of our results, we show that for octal quantizers, the low-SNR asymptotic GMI is higher than the known lower bound of channel capacity in the literature [4].

Finally, we explore the benefit of super-Nyquist sampling. Considering a real Gaussian channel with a bandlimited pulse-shaping function and with general memoryless output distortion, we obtain a formula for its GMI, when the channel output is uniformly sampled at  $L$  times the Nyquist rate. We then particularize to the case of binary symmetric output quantization. We demonstrate through numerical evaluation that super-Nyquist sampling leads to benefit in terms of increased GMI over all SNR, for different values of  $L$ . In the low-SNR regime, the asymptotic GMI we obtain for  $L = 2$  with a carefully chosen pulse-shaping function almost coincides with the known lower bound of channel capacity in the literature [7]. In the high-SNR regime, we make an interesting observation that, when the sampling rate is sufficiently high, the GMI becomes greater than one bit/c.u.. At first glance, this result is surprising since the output quantization is binary; however, it is in fact reasonable, because for each channel input symbol, there are multiple binary output symbols due to super-Nyquist sampling, and the amount of information carried by the Gaussian codebook ensemble exceeds one bit per input symbol.

We organize the remaining part of the paper as follows. Section II describes the general Nyquist-sampled channel model and establishes the general GMI formula. Section III treats the scenario where only transmit-side distortion exists, revisiting the well-known decomposition of Bussgang's theorem. Section IV treats the channel model with binary symmetric output quantization. Section V treats symmetric output quantizers with more than two quantization levels. Section VI explores the benefit of super-Nyquist sampling. Finally Section VII concludes the paper. Auxiliary technical derivations and other supporting results are archived in the Supplementary Material.

## II. GENERAL FRAMEWORK FOR REAL-VALUED NYQUIST-SAMPLED CHANNELS

With Nyquist sampling, it loses no generality to consider a discrete-time channel model, with a sequence  $\{Z_k\}$  of independent and identically distributed (i.i.d.) real Gaussian noise, *i.e.*,  $Z_k \sim \mathcal{N}(0, \sigma^2)$ . The channel input symbols constitute a sequence  $\{X_k\}$ . Without distortion, the

received signal is  $Y. = X. + Z.$ . However, the distortion may affect both the channel input and the channel output. A memoryless distortion, in general form, is a deterministic mapping  $f(\cdot)$ , which transforms a pair of channel input symbol and noise sample  $(x, z)$  into a real number  $f(x, z)$ . Hence the channel observation at the decoder is

$$W_k = f(X_k, Z_k), \quad \text{for } k = 1, 2, \dots, n, \quad (1)$$

where  $n$  denotes the codeword length; see the illustration in Figure 1. We note that, such a form of distortion can describe the case where the channel output  $Y. = X. + Z.$  is distorted, *i.e.*,  $w = f(x, z) = f_o(x + z)$ , or the case where the channel input  $X.$  is distorted by the transmitter, *i.e.*,  $w = f(x, z) = f_i(x) + z$ , or the case where both input and output are distorted, *i.e.*,  $w = f(x, z) = f_o(f_i(x) + z)$ .

For transmission, the source selects a message  $M$  from  $\mathcal{M} = \{1, 2, \dots, \lfloor e^{nR} \rfloor\}$  uniformly randomly, and maps the selected message to a transmitted codeword, which is a length- $n$  real sequence,  $\{X_k(M)\}_{k=1}^n$ . We restrict the codebook to be an i.i.d.  $\mathcal{N}(0, \mathcal{E}_s)$  ensemble. That is, each codeword is a sequence of  $n$  i.i.d.  $\mathcal{N}(0, \mathcal{E}_s)$  random variables, and all the codewords are mutually independent. Such a choice of codebook ensemble satisfies the average power constraint  $\frac{1}{n} \sum_{k=1}^n \mathbf{E} X_k^2(M) \leq \mathcal{E}_s$ . We thus define the nominal SNR as  $\text{SNR} = \mathcal{E}_s / \sigma^2$ .

As is well known, when transceiver distortion is absent (*i.e.*,  $w = y$ ), as the codeword length  $n$  grows without bound, the Gaussian codebook ensemble achieves the capacity of the channel,  $\frac{1}{2} \log(1 + \text{SNR})$ . In the following, we proceed to investigate the GMI when the channel experiences the memoryless nonlinear distortion  $f(\cdot)$ .

To proceed, we restrict the decoder to follow a nearest-neighbor rule, which, upon observing  $\{w_k\}_{k=1}^n$ , computes for all possible messages, the distance metric,

$$D(m) = \frac{1}{n} \sum_{k=1}^n [w_k - ax_k(m)]^2, \quad m \in \mathcal{M}, \quad (2)$$

and decides the received message as  $\hat{m} = \arg \min_{m \in \mathcal{M}} D(m)$ . In (2), the parameter  $a$  is selected appropriately for optimizing the decoding performance. We note that, the nearest-neighbor decoder (with  $a = 1$ ) coincides with the optimal (maximum-likelihood) decoder in the absence of distortion, but is in general suboptimal (mismatched) for the distorted channel (1).

In the subsequent development in this section, we characterize an achievable rate which guarantees that the average probability of decoding error decreases to zero as  $n \rightarrow \infty$ , for

Gaussian codebook ensemble and the nearest-neighbor decoding rule, following the argument used in [10]. When we consider the average probability of decoding error averaged over both the message set and the Gaussian codebook ensemble, due to the symmetry in the codebook, it suffices to condition upon the scenario where the message  $m = 1$  is selected for transmission.

With  $m = 1$ , we have

$$\begin{aligned} \lim_{n \rightarrow \infty} D(1) &= \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{k=1}^n [W_k - aX_k(1)]^2 = \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{k=1}^n [f(X_k, Z_k) - aX_k(1)]^2 \\ &= \mathbf{E} \{ [f(X, Z) - aX]^2 \} \quad \text{a.s.} \end{aligned} \quad (3)$$

where  $X \sim \mathcal{N}(0, \mathcal{E}_s)$  and  $Z \sim \mathcal{N}(0, \sigma^2)$ , from the law of large numbers.

The exponent of the probability of decoding error is the GMI, given by

$$I_{\text{GMI}} = \sup_{\theta < 0} (\theta \mathbf{E} \{ [f(X, Z) - aX]^2 \} - \Lambda(\theta)), \quad (4)$$

where

$$\Lambda(\theta) = \lim_{n \rightarrow \infty} \frac{1}{n} \Lambda_n(n\theta), \quad (5)$$

$$\Lambda_n(n\theta) = \log \mathbf{E} \{ e^{n\theta D(m)} | W_k, k = 1, \dots, n \}, \quad \forall m \neq 1. \quad (6)$$

From Chernoff's bound and the union upper bounding technique, we see that as long as the information rate is less than  $I_{\text{GMI}}$ , the average probability of decoding error decreases to zero as  $n \rightarrow \infty$ . Therefore, the GMI serves as a reasonable lower bound for the achievable information rate for a given codebook ensemble and a given decoding rule.

After the mathematical manipulation given in Supplementary Material VII-A, we establish the following result.

*Proposition 1:* With Gaussian codebook ensemble and nearest-neighbor decoding, the GMI of the distorted channel (1) is

$$I_{\text{GMI}} = \frac{1}{2} \log \left( 1 + \frac{\Delta}{1 - \Delta} \right), \quad (7)$$

where the parameter  $\Delta$  is

$$\Delta = \frac{\{\mathbf{E}[f(X, Z)X]\}^2}{\mathcal{E}_s \mathbf{E}[f(X, Z)]^2}. \quad (8)$$

The corresponding optimal choice of the decoding scaling parameter  $a$  is  $a_{\text{opt}} = \mathbf{E}[f(X, Z)X] / \mathcal{E}_s$ .

We readily see that  $\Delta$  is the squared correlated coefficient between the channel input  $X$  and the distorted channel output  $f(X, Z)$ , which is upper bounded by one, from Cauchy-Schwartz inequality. A larger value of  $\Delta$  corresponds to a higher effective SNR.

When contrasted with the capacity of the undistorted channel,  $\frac{1}{2} \log(1 + \text{SNR})$ , we can define the effective SNR of the distorted channel as  $\text{SNR}_e = \frac{\Delta}{1-\Delta}$ .

As an immediate verification, consider the undistorted channel  $W = X + Z$ , for which we have  $\Delta = \mathcal{E}_s / (\mathcal{E}_s + \sigma^2)$ . Consequently, the effective SNR is  $\text{SNR}_e = \mathcal{E}_s / \sigma^2$ , leading to the capacity of the undistorted channel.

It is perhaps worth noting that, the derivation of the GMI in fact does not require  $Z$  be Gaussian. Indeed, as long as  $\{Z_k\}$  is an ergodic process and is independent of  $\{X_k\}$ , the general result of Proposition 1 holds. However, for simplicity, in the current paper we confine ourselves to i.i.d. Gaussian noise, and do not pursue this issue further.

*Remark on Antipodal Codebook Ensemble:* The foregoing analysis of GMI applies to any input distribution. Here, consider antipodal inputs, *i.e.*,  $X_k(m)$  takes  $\sqrt{\mathcal{E}_s}$  and  $-\sqrt{\mathcal{E}_s}$  with probability  $1/2$ , respectively. All the codeword symbols are mutually independent. Again, we consider a nearest-neighbor decoding rule, with distance metric in form of (2). Following the same line of analysis as that for the Gaussian codebook ensemble, we have

$$I_{\text{GMI}} = \sup_{t \in \mathbb{R}} \left( t \mathbf{E}[X f(X, Z)] - \mathbf{E} \log \cosh(t \sqrt{\mathcal{E}_s} f(X, Z)) \right), \quad (9)$$

and the optimal value of  $t$  should satisfy

$$\mathbf{E} \left[ \sqrt{\mathcal{E}_s} f(X, Z) \cdot \tanh(t \sqrt{\mathcal{E}_s} f(X, Z)) \right] = \mathbf{E}[X f(X, Z)]. \quad (10)$$

Supplementary Material VII-B. The evaluation of the GMI is usually more difficult than that for the Gaussian codebook ensemble.

### III. CHANNELS WITH TRANSMIT-SIDE DISTORTION: BUSSGANG REVISITED

In this section, we briefly consider the scenario where only the channel input is distorted, *i.e.*,  $w = f_i(x) + z$ . Since  $X$  and  $Z$  are independent, the optimal choice of the decoding scaling parameter becomes

$$a_{\text{opt}} = \frac{\mathbf{E}[(f_i(X) + Z)X]}{\mathcal{E}_s} = \frac{\mathbf{E}[X f_i(X)]}{\mathcal{E}_s}. \quad (11)$$



The resulting value of  $\Delta$  is

$$\Delta = \frac{\{\mathbf{E}[Xf_i(X)]\}^2}{\mathcal{E}_s (\mathbf{E}[f_i(X)]^2 + \sigma^2)}, \quad (12)$$

and the effective SNR is

$$\text{SNR}_e = \frac{\Delta}{1 - \Delta} = \frac{\{\mathbf{E}[Xf_i(X)]\}^2}{\mathcal{E}_s (\mathbf{E}[f_i(X)]^2 + \sigma^2) - \{\mathbf{E}[Xf_i(X)]\}^2}. \quad (13)$$

Inspecting  $a_{\text{opt}}$  in (11), we notice that it leads to the following decomposition of  $f_i(X)$ :

$$f_i(X) = a_{\text{opt}}X + V, \quad (14)$$

where the distortion  $V$  is uncorrelated with the input  $X$ . Recalling the Bussgang decomposition [2], we conclude that, when there is only transmit-side distortion, the optimal decoding scaling parameter in the nearest-neighbor decoding rule coincides with that suggested by Bussgang's theorem. Note that this conclusion does not hold in general when receive-side distortion exists.

#### IV. CHANNELS WITH BINARY SYMMETRIC OUTPUT QUANTIZATION

In this section, we consider the scenario where the channel output  $Y = X + Z$  passes through a binary symmetric hard-limiter to retain its sign information only. This is also called one-bit/mono-bit quantization/analog-to-digital conversion, and we can write it as  $w = f(x, z) = \text{sgn}(x + z)$ .

For this scenario, we have

$$\Delta = \frac{\{\mathbf{E}[X \cdot \text{sgn}(X + Z)]\}^2}{\mathcal{E}_s}, \quad (15)$$

where we use the fact that the average output power  $\mathbf{E}[\text{sgn}(X + Z)]^2$  is unity. Now in order to facilitate the evaluation of the expectation in the numerator in (15), we introduce the “partial mean” of the random variable  $X \sim \mathcal{N}(0, \mathcal{E}_s)$

$$F(z) = \int_z^\infty \frac{x}{\sqrt{2\pi\mathcal{E}_s}} e^{-\frac{x^2}{2\mathcal{E}_s}} dx = \sqrt{\frac{\mathcal{E}_s}{2\pi}} \exp\left(-\frac{z^2}{2\mathcal{E}_s}\right), \quad (16)$$

which is an even function of  $z \in (-\infty, \infty)$ . We denote by  $p_X(x)$  and  $p_Z(z)$  the probability density functions of  $X \sim \mathcal{N}(0, \mathcal{E}_s)$  and  $Z \sim \mathcal{N}(0, \sigma^2)$ , respectively, and proceed as

$$\begin{aligned} \mathbf{E}[X \cdot \text{sgn}(X + Z)] &= \iint_{x+z>0} xp_X(x)p_Z(z)dx dz - \iint_{x+z<0} xp_X(x)p_Z(z)dx dz \\ &= 2 \iint_{x+z>0} xp_X(x)p_Z(z)dx dz = 2 \int_{-\infty}^{\infty} p_Z(z)F(-z)dz = \mathcal{E}_s \sqrt{\frac{2}{\pi(\mathcal{E}_s + \sigma^2)}}. \end{aligned} \quad (17)$$

This leads to

$$\Delta = \frac{\mathcal{E}_s^2 \frac{2}{\pi(\mathcal{E}_s + \sigma^2)}}{\mathcal{E}_s} = \frac{2\mathcal{E}_s}{\pi(\mathcal{E}_s + \sigma^2)}, \quad (18)$$

and

$$\text{SNR}_e = \frac{\Delta}{1 - \Delta} = \frac{2\mathcal{E}_s}{(\pi - 2)\mathcal{E}_s + \pi\sigma^2}. \quad (19)$$

So we get the following asymptotic behavior:

- High SNR: When  $\text{SNR} = \mathcal{E}_s/\sigma^2 \rightarrow \infty$ ,

$$\text{SNR}_e = \frac{2}{\pi - 2} - \frac{2\pi}{(\pi - 2)^2} \frac{1}{\text{SNR}} + o\left(\frac{1}{\text{SNR}}\right), \quad (20)$$

$$I_{\text{GMI}} = \frac{1}{2} \log \frac{\pi}{\pi - 2} - \frac{1}{\pi - 2} \frac{1}{\text{SNR}} + o\left(\frac{1}{\text{SNR}}\right). \quad (21)$$

- Low SNR: When  $\text{SNR} \rightarrow 0$ ,

$$\text{SNR}_e = \frac{2}{\pi} \text{SNR} - \frac{2(\pi - 2)}{\pi^2} \text{SNR}^2 + o(\text{SNR}^2), \quad (22)$$

$$I_{\text{GMI}} = \frac{1}{\pi} \text{SNR} - \frac{\pi - 1}{\pi^2} \text{SNR}^2 + o(\text{SNR}^2). \quad (23)$$

We make two observations. First, at high SNR, the GMI converges to 0.7302 bits/c.u., strictly less than the limit of the channel capacity, 1 bit/c.u., revealing that the suboptimal Gaussian codebook ensemble leads to non-negligible penalty when the effect of distortion is dominant. Second, at low SNR, the ratio between the GMI and the SNR converges to  $1/\pi$ , and thus asymptotically coincides with the behavior of the channel capacity [4]. Intuitively, this is because in the low-SNR regime the effect of noise is dominant, and thus the channel is approximately still Gaussian. In Figure 2 we plot the GMI  $I_{\text{GMI}}$  and the channel capacity  $C = 1 - H_2(Q(\sqrt{\mathcal{E}_s/\sigma^2}))$  [5] versus SNR. The different behaviors of the GMI in the two regimes are evident in the figure.

## V. CHANNELS WITH MULTI-BIT OUTPUT QUANTIZATION

In this section, we continue the exploration of output quantization and consider specifically the scenario where the channel output  $Y$  passes through a  $2M$ -level symmetric quantizer, as

$$w = f(x + z) = r_i \cdot \text{sgn}(x + z) \quad \text{if } |x + z| \in [\alpha_{i-1}, \alpha_i), \quad (24)$$

for  $i = 1, \dots, M$ , where  $\alpha_0 = 0 < \alpha_1 < \dots < \alpha_M = \infty$ . The parameters include the reconstruction points  $\{r_1, \dots, r_M\}$ , and the quantization thresholds  $\{\alpha_1, \dots, \alpha_{M-1}\}$ . Note that with  $2M$  levels, the quantizer bit-width is  $(\log_2 M + 1)$  bits.

For a  $2M$ -level symmetric quantizer, we can evaluate that (see Supplementary Material VII-D)

$$\mathbf{E}[f(X+Z)]^2 = 2 \sum_{i=1}^M r_i^2 \left[ Q\left(\frac{\alpha_{i-1}}{\sqrt{\mathcal{E}_s + \sigma^2}}\right) - Q\left(\frac{\alpha_i}{\sqrt{\mathcal{E}_s + \sigma^2}}\right) \right], \quad (25)$$

where the Q-function is  $Q(z) = \frac{1}{\sqrt{2\pi}} \int_z^\infty e^{-x^2/2} dx$ , and

$$\mathbf{E}[f(X+Z)X] = \mathcal{E}_s \sqrt{\frac{2}{\pi(\mathcal{E}_s + \sigma^2)}} \sum_{i=1}^M r_i \left[ e^{-\frac{\alpha_{i-1}^2}{2(\mathcal{E}_s + \sigma^2)}} - e^{-\frac{\alpha_i^2}{2(\mathcal{E}_s + \sigma^2)}} \right]. \quad (26)$$

To further simplify the notation, define  $\tilde{Q}(z) = \frac{1}{2\sqrt{\pi}} \int_0^z (-\log x)^{-1/2} dx$  for  $z \in [0, 1]$ ,<sup>3</sup> and introduce  $t_i = e^{-\frac{\alpha_i^2}{2(\mathcal{E}_s + \sigma^2)}}$  for  $i = 0, 1, \dots, M$  with  $t_0 = 1 > t_1 > \dots > t_M = 0$ . We thus can rewrite

$$\mathbf{E}[f(X+Z)]^2 = 2 \sum_{i=1}^M r_i^2 [\tilde{Q}(t_{i-1}) - \tilde{Q}(t_i)], \quad (27)$$

$$\mathbf{E}[f(X+Z)X] = \mathcal{E}_s \sqrt{\frac{2}{\pi(\mathcal{E}_s + \sigma^2)}} \sum_{i=1}^M r_i (t_{i-1} - t_i). \quad (28)$$

These lead to

$$\Delta = \frac{\mathcal{E}_s}{\pi(\mathcal{E}_s + \sigma^2)} \frac{\left[ \sum_{i=1}^M r_i (t_{i-1} - t_i) \right]^2}{\sum_{i=1}^M r_i^2 [\tilde{Q}(t_{i-1}) - \tilde{Q}(t_i)]}. \quad (29)$$

In (29), the second term is independent of the SNR, and can be optimized separately. Let us denote this term by  $K_{\underline{r}, \underline{t}}$ , and write  $\Delta = \frac{\mathcal{E}_s K_{\underline{r}, \underline{t}}}{\pi(\mathcal{E}_s + \sigma^2)}$ . We consequently have the following effective SNR:

$$\text{SNR}_e = \frac{K_{\underline{r}, \underline{t}} \mathcal{E}_s}{(\pi - K_{\underline{r}, \underline{t}}) \mathcal{E}_s + \pi \sigma^2}. \quad (30)$$

- High SNR: When  $\text{SNR} \rightarrow \infty$ ,

$$\text{SNR}_e = \frac{K_{\underline{r}, \underline{t}}}{\pi - K_{\underline{r}, \underline{t}}} - \frac{K_{\underline{r}, \underline{t}} \pi}{(\pi - K_{\underline{r}, \underline{t}})^2} \frac{1}{\text{SNR}} + o\left(\frac{1}{\text{SNR}}\right), \quad (31)$$

$$I_{\text{GMI}} = \frac{1}{2} \log \frac{\pi}{\pi - K_{\underline{r}, \underline{t}}} - \frac{K_{\underline{r}, \underline{t}}}{2(\pi - K_{\underline{r}, \underline{t}})} \frac{1}{\text{SNR}} + o\left(\frac{1}{\text{SNR}}\right). \quad (32)$$

- Low SNR: When  $\text{SNR} \rightarrow 0$ ,

$$\text{SNR}_e = \frac{K_{\underline{r}, \underline{t}}}{\pi} \text{SNR} - \frac{K_{\underline{r}, \underline{t}}(\pi - K_{\underline{r}, \underline{t}})}{\pi^2} \text{SNR}^2 + o(\text{SNR}^2), \quad (33)$$

$$I_{\text{GMI}} = \frac{K_{\underline{r}, \underline{t}}}{2\pi} \text{SNR} - \frac{K_{\underline{r}, \underline{t}}(\pi - K_{\underline{r}, \underline{t}}/2)}{2\pi^2} \text{SNR}^2 + o(\text{SNR}^2). \quad (34)$$

<sup>3</sup>We have  $\tilde{Q}(z) = Q(\sqrt{-2 \log z}) = (1/2) \cdot \text{erfc}(\sqrt{-\log z})$ .

It is thus apparent that the value of  $K_{\underline{r}, \underline{t}}$  determines the system performance, for all SNR. We hence seek to maximize

$$K_{\underline{r}, \underline{t}} = \frac{\left[ \sum_{i=1}^M r_i (t_{i-1} - t_i) \right]^2}{\sum_{i=1}^M r_i^2 [\tilde{Q}(t_{i-1}) - \tilde{Q}(t_i)]}, \quad (35)$$

where  $t_0 = 1 > t_1 > \dots > t_M = 0$  and  $r_i \geq 0$  for all  $i = 1, \dots, M$ .

Taking the partial derivatives of  $K_{\underline{r}, \underline{t}}$  with respect to  $r_i$ ,  $i = 1, \dots, M$ , and enforcing them to vanish, we have that the following set of equations needs to hold for maximizing  $K_{\underline{r}, \underline{t}}$ ,

$$r_i = \frac{t_{i-1} - t_i}{\tilde{Q}(t_{i-1}) - \tilde{Q}(t_i)} \frac{\sum_{j=1}^M r_j^2 [\tilde{Q}(t_{j-1}) - \tilde{Q}(t_j)]}{\sum_{j=1}^M r_j (t_{j-1} - t_j)}, \quad i = 1, \dots, M. \quad (36)$$

Substituting these  $\{r_i\}$  into  $K_{\underline{r}, \underline{t}}$  and simplifying the resulting expression, we obtain

$$K_{\underline{t}} = \max_{\underline{r}} K_{\underline{r}, \underline{t}} = \sum_{i=1}^M \frac{(t_{i-1} - t_i)^2}{\tilde{Q}(t_{i-1}) - \tilde{Q}(t_i)}. \quad (37)$$

That is, the optimal quantizer design should solve the following maximization problem:

$$\max_{\underline{t}} \sum_{i=1}^M \frac{(t_{i-1} - t_i)^2}{\tilde{Q}(t_{i-1}) - \tilde{Q}(t_i)}, \quad \text{s.t.} \quad t_0 = 1 > t_1 > \dots > t_M = 0. \quad (38)$$

*Example:* Fine quantization,  $\max_{i=1, \dots, M} (t_{i-1} - t_i) \rightarrow 0$

In this case, the following approximation becomes accurate:

$$\frac{\tilde{Q}(t_{i-1}) - \tilde{Q}(t_i)}{t_{i-1} - t_i} \approx \tilde{Q}'(t_{i-1}), \quad \forall i = 1, \dots, M. \quad (39)$$

So the resulting  $K_{\underline{t}}$  behaves like

$$\begin{aligned} K_{\underline{t}} &= \sum_{i=1}^M \frac{(t_{i-1} - t_i)^2}{\tilde{Q}(t_{i-1}) - \tilde{Q}(t_i)} \rightarrow \int_0^1 \frac{1}{\tilde{Q}'(t)} dt \\ &= 2\sqrt{\pi} \int_0^1 \sqrt{-\log t} dt = 2\sqrt{\pi} \int_{-\infty}^{\infty} y^2 e^{-y^2} dy = \pi. \end{aligned} \quad (40)$$

Therefore, as the quantization goes fine asymptotically, the effective SNR as given by (30) approaches the actual SNR, and thus the performance loss due to quantization eventually diminishes.

*Example:* 4-level quantization,  $M = 2$

In this case, there is only one variable,  $t = t_1$ , to optimize. The maximization problem becomes

$$\max_{t \in (0,1)} \frac{(1-t)^2}{1/2 - \tilde{Q}(t)} + \frac{t^2}{\tilde{Q}(t)}. \quad (41)$$

A numerical computation immediately gives  $\max_{t \in (0,1)} K_t = 2.7775$ , and interestingly, the maximizing  $t = 0.618$  is the golden ratio.

*Example: Uniform quantization*

In practical systems, uniform quantization is common, in which the thresholds satisfy  $\alpha_i = i\sqrt{2(\mathcal{E}_s + \sigma^2)\alpha}$  for  $i = 0, 1, \dots, M-1$ , and  $\alpha_M = \infty$ , where  $\alpha > 0$  is a parameter for optimization. These thresholds lead to

$$K_{\underline{t}} = \sum_{i=1}^{M-1} \frac{\left[ e^{-(i-1)^2\alpha} - e^{-i^2\alpha} \right]^2}{Q(\sqrt{2\alpha}(i-1)) - Q(\sqrt{2\alpha}i)} + \frac{e^{-2(M-1)^2\alpha}}{Q(\sqrt{2\alpha}(M-1))}, \quad (42)$$

which can be further maximized over  $\alpha > 0$ .

In Table I, we list the numerical results for optimizing  $K_{\underline{t}}$  over  $\alpha$ , up until  $M = 8$ .

*Example:  $t$ -uniform quantization*

An alternative quantizer design is to let the values of  $\underline{t}$  be uniformly placed within  $[0, 1]$ , *i.e.*,  $t_i = (M-i)/M$ , for  $i = 0, 1, \dots, M$ . This quantization leads to

$$K_{\underline{t}} = \frac{1}{M^2} \sum_{i=1}^M \frac{1}{\tilde{Q}(1 - (i-1)/M) - \tilde{Q}(1 - i/M)}. \quad (43)$$

In Table II, we list the numerical results of  $K_{\underline{t}}$  for  $t$ -uniform quantizers, up until  $M = 8$ . We notice that the  $t$ -uniform quantization is consistently inferior to the optimized uniform quantization.

*Example: Optimal quantization*

We can also develop program to numerically solve the optimization problem (38). In Table III, we list the results, up until  $M = 8$ . We also list the value of the optimal  $t_1$ , from which we can recursively compute the whole optimal  $\underline{t}$  vector, through enforcing the partial derivatives  $\partial K_{\underline{t}}/\partial t_i$  to vanish for  $i = 2, \dots, M-1$  progressively.

From the numerical results in the above examples, we observe that the GMI may be fairly close to the channel capacity at low SNR. For example, with the optimal octal quantizer ( $M = 4$ ), the low-SNR GMI scales with SNR like  $0.4827 \cdot \text{SNR}$  bits/c.u., which is better than the known lower bound  $0.475 \cdot \text{SNR}$  bits/c.u. in the literature [4]. In Figure 3 we plot the GMI  $I_{\text{GMI}}$  achieved by the optimal quantizers, for  $M = 2, 3, \dots, 8$ . For comparison we also plot in dash-dot curve the capacity  $(1/2)\log_2(1 + \text{SNR})$  of undistorted channels. We can roughly assess that, with  $M = 4$  (*i.e.*, 3-bit quantization), the performance gap between the GMI and the undistorted

channel capacity is mild up until  $\text{SNR} \approx 10$  dB; and with  $M = 8$  (i.e., 4-bit quantization), the performance gap is mild up until  $\text{SNR} \approx 15$  dB. Compared with the numerically evaluated capacity for 2/3-bit quantization in [5], we see that using the Gaussian codebook ensemble and the nearest-neighbor decoding rule induce a 15-25% rate loss at high SNR. Comparing Tables I and III, we further notice that the performance loss due to using uniform quantization is essentially negligible.

*Remark on Possible Connection with Capacity per Unit Cost:* For a given  $2M$ -level symmetric quantizer, we can evaluate the channel capacity per unit cost (symbol energy in our context) by optimizing a single nonzero input symbol,  $x$  (see [12]). Without loss of generality, we let  $x > 0$  and the noise variance  $\sigma^2$  be unity. Then the capacity per unit cost can be evaluated as

$$\sup_{x>0} \frac{1}{x^2} \sum_{i=1}^M \left[ (Q(\alpha_{i-1} - x) - Q(\alpha_i - x)) \log \frac{Q(\alpha_{i-1} - x) - Q(\alpha_i - x)}{Q(\alpha_{i-1}) - Q(\alpha_i)} \right. \\ \left. + (Q(\alpha_{i-1} + x) - Q(\alpha_i + x)) \log \frac{Q(\alpha_{i-1} + x) - Q(\alpha_i + x)}{Q(\alpha_{i-1}) - Q(\alpha_i)} \right]. \quad (44)$$

With some manipulations, we find that  $K_t/(2\pi)$  is exactly the limit value of the term in (44) as  $x \rightarrow 0$ .<sup>4</sup> Therefore, only if the capacity per unit cost (44) is achieved by  $x \rightarrow 0$ , the GMI coincides with the channel capacity in the low-SNR limit. Unfortunately, as revealed by our numerical experiments, this condition does not generally hold for all possible symmetric quantizers.

## VI. SUPER-NYQUIST OUTPUT SAMPLING

In this section, we examine the scenario where we sample the channel output at a rate higher than the Nyquist rate, and investigate the benefit of increased sampling rates in terms of the GMI.

We start with a continuous-time baseband model in which the transmitted signal is

$$x(t) = \frac{1}{\sqrt{2W}} \sum_{k=1}^n X_k g\left(t - \frac{k}{2W}\right), \quad (45)$$

where  $g(\cdot)$  is a pulse function with unit energy and is band limited within  $W$  Hz. In analysis, a commonly used pulse function is the sinc function  $g(t) = \sqrt{2W} \text{sinc}(2Wt)$  with  $\text{sinc}(t) = \sin(\pi t)/(\pi t)$ , which vanishes at the Nyquist sampling time instants  $t = \{k/(2W)\}_{k=-\infty}^{\infty}$ . The

<sup>4</sup>This is also half of the Fisher information for estimating  $X = 0$  from the quantized channel output  $W$  [12].

channel input is a sequence of independent  $\mathcal{N}(0, \mathcal{E}_s)$  random variables  $\{X_k\}_{k=1}^n$ . With additive white Gaussian noise  $z(t)$ , the received signal is

$$y(t) = x(t) + z(t). \quad (46)$$

We assume that  $z(t)$  is band-limited within  $W$  Hz, with in-band two-sided power spectral density  $\sigma^2/2$ . So the autocorrelation function of  $z(t)$  is  $K_z(\tau) = \frac{\sigma^2}{2} \text{sinc}(2W\tau)$ .

We consider a uniform sampler, which samples the channel output  $y(t)$  at  $L$  times the Nyquist rate. For the  $k$ -th input symbol, the sampling time instants thus are

$$t = \left\{ \frac{k}{2W} + \frac{l}{2WL} - \tau_L \right\}_{l=0}^{2(L-1)}. \quad (47)$$

Here,  $\tau_L$  is a constant offset to ensure that the sampling times are symmetric with respect to the center of the  $k$ -th input symbol pulse; for example,  $\tau_1 = 0$  (Nyquist sampling),  $\tau_2 = 1/(4W)$ ,  $\tau_3 = 1/(3W)$ ,  $\tau_4 = 3/(8W)$ ... Generally,  $\tau_L = \frac{L-1}{L} \frac{1}{2W}$ . Thus we can rewrite (47) as

$$t = \frac{1}{2W} \left\{ k + \frac{l}{L} \right\}_{l=-L+1}^{L-1}. \quad (48)$$

Denote the output samples by  $\{Y_{k,l}\}$  with  $Y_{k,l} = y(t_{k,l})$  where  $t_{k,l} = \frac{1}{2W}(k + l/L)$ . The samples pass through a nonlinear distortion device, so that the observed samples are  $W_{k,l} = f(Y_{k,l})$ .

Let us generalize the nearest-neighbor decoding rule in Section II as follows. For all possible messages, the decoder computes the distance metrics,

$$D(m) = \frac{1}{n} \sum_{k=1}^n \sum_{l=-L+1}^{L-1} \xi_l [w_{k,l} - a_l x_k(m)]^2, \quad m \in \mathcal{M}, \quad (49)$$

where  $\{\xi_l\}_{l=-L+1}^{L-1}$  and  $\{a_l\}_{l=-L+1}^{L-1}$  are weighting coefficients, and decides the received message as  $\hat{m} = \arg \min_{m \in \mathcal{M}} D(m)$ . We then note that

$$\begin{aligned} \sum_{l=-L+1}^{L-1} \xi_l [w_{k,l} - a_l x_k(m)]^2 &= \sum_{l=-L+1}^{L-1} \xi_l w_{k,l}^2 - 2x_k(m) \sum_{l=-L+1}^{L-1} \xi_l a_l w_{k,l} + x_k^2(m) \sum_{l=-L+1}^{L-1} \xi_l a_l^2 \\ &= \left( \sum_{l=-L+1}^{L-1} \xi_l a_l^2 \right) \cdot \left[ x_k(m) - \frac{\sum_{l=-L+1}^{L-1} \xi_l a_l w_{k,l}}{\sum_{l=-L+1}^{L-1} \xi_l a_l^2} \right]^2 + \left[ \sum_{l=-L+1}^{L-1} \xi_l w_{k,l}^2 - \frac{\left( \sum_{l=-L+1}^{L-1} \xi_l a_l w_{k,l} \right)^2}{\sum_{l=-L+1}^{L-1} \xi_l a_l^2} \right]. \end{aligned}$$

Therefore, without loss of generality, we may consider the simplified nearest-neighbor decoding distance metric

$$D(m) = \frac{1}{n} \sum_{k=1}^n \left[ \sum_{l=-L+1}^{L-1} \beta_l w_{k,l} - x_k(m) \right]^2, \quad (50)$$

for which the tunable weighting coefficients are  $\underline{\beta} = \{\beta_l\}_{l=-L+1}^{L-1}$ .

Following the same procedure as that in Section II for the Nyquist-sampled channel model, we first examine the limit of  $D(1)$  assuming that the message  $m = 1$  is sent. Since the channel input symbols  $X_k$  are i.i.d. and the noise process is wide-sense stationary, the observed samples  $W_{k,l}$  constitute an ergodic process.<sup>5</sup> Consequently, we have

$$\lim_{n \rightarrow \infty} D(1) = \mathbf{E} \left[ \sum_{l=-L+1}^{L-1} \beta_l W_{0,l} - X_0 \right]^2 \quad \text{a.s.} \quad (51)$$

On the other hand, for any  $m \neq 1$ , we have

$$\begin{aligned} \frac{1}{n} \Lambda_n(n\theta) &= \frac{1}{n} \log \mathbf{E} \left\{ e^{\theta \sum_{k=1}^n [\sum_{l=-L+1}^{L-1} \beta_l W_{k,l} - X_k(m)]^2} \middle| W_{k,l}, k = 1, \dots, n, l = -L+1, \dots, L-1 \right\} \\ &= \frac{\theta}{1 - 2\theta \mathcal{E}_s} \frac{1}{n} \sum_{k=1}^n \left[ \sum_{l=-L+1}^{L-1} \beta_l W_{k,l} \right]^2 - \frac{1}{2} \log(1 - 2\theta \mathcal{E}_s) \\ &\rightarrow \frac{\theta}{1 - 2\theta \mathcal{E}_s} \mathbf{E} \left[ \sum_{l=-L+1}^{L-1} \beta_l W_{0,l} \right]^2 - \frac{1}{2} \log(1 - 2\theta \mathcal{E}_s) \quad \text{a.s.} \end{aligned} \quad (52)$$

In both limits above,  $\{W_{0,l}\}_{l=-L+1}^{L-1}$  are induced by an infinite sequence of inputs,  $\{X_k\}_{k=-\infty}^{\infty}$ .

So the GMI is

$$I_{\text{GMI}} = \sup_{\underline{\beta}, \theta < 0} \left\{ \theta \mathbf{E} \left[ \sum_{l=-L+1}^{L-1} \beta_l W_{0,l} - X_0 \right]^2 - \frac{\theta}{1 - 2\theta \mathcal{E}_s} \mathbf{E} \left[ \sum_{l=-L+1}^{L-1} \beta_l W_{0,l} \right]^2 + \frac{1}{2} \log(1 - 2\theta \mathcal{E}_s) \right\}, \quad (53)$$

and we have the following result, whose derivation is given in Supplementary Material VII-G.

*Proposition 2:* The GMI with super-Nyquist output sampling is

$$I_{\text{GMI}} = \frac{1}{2} \log \left( 1 + \frac{\Delta}{1 - \Delta} \right), \quad (54)$$

where  $\Delta = (\underline{b}^T \mathbf{\Omega}^{-1} \underline{b}) / \mathcal{E}_s$ ,  $\mathbf{\Omega}$  is a  $(2L - 1) \times (2L - 1)$  matrix with its  $(u, l)$ -element being  $\mathbf{E}[W_{0,u} W_{0,l}]$ , and  $\underline{b}$  is a  $(2L - 1)$ -dimensional vector with its  $l$ -element being  $\mathbf{E}[X_0 W_{0,l}]$ ,  $u, l = -L + 1, \dots, L - 1$ . To achieve the GMI in (54), the optimal weighting coefficients are

$$\underline{\beta} = \frac{\mathcal{E}_s}{\underline{b}^T \mathbf{\Omega}^{-1} \underline{b}} \mathbf{\Omega}^{-1} \underline{b}. \quad (55)$$

We notice that the GMI in (54) is a natural extension of that in Proposition 1 for the Nyquist-sampling case, and we can also define the effective SNR by  $\text{SNR}_e = \Delta / (1 - \Delta)$ .

<sup>5</sup>We note that the transmission of a codeword,  $\{X_k\}_{k=1}^n$ , is finite-length. In order to meet the ergodicity condition, we may slightly modify the model by appending  $\{X_k\}_{k=-\infty}^0$  and  $\{X_k\}_{k=n+1}^{\infty}$ , which consist of i.i.d.  $\mathcal{N}(0, \mathcal{E}_s)$  random variables as additional interference, to the transmitted codeword.



### A. Binary Symmetric Quantization: Sinc Pulse Function

We examine binary symmetric quantization in which  $w = \text{sgn}(y)$ . For this purpose, we need to evaluate  $\Omega$  and  $\underline{b}$ . For each  $l$ ,

$$Y_{0,l} = \frac{1}{\sqrt{2W}} \sum_{k=-\infty}^{\infty} X_k g\left(\frac{l}{2WL} - \frac{k}{2W}\right) + Z\left(\frac{l}{2WL}\right). \quad (56)$$

Utilizing (17) and noting that  $\{X_k\}$  are i.i.d., we have

$$\begin{aligned} b_l &= \mathbf{E}[X_0 \text{sgn}(Y_{0,l})] \\ &= \frac{\mathcal{E}_s g(l/(2WL))}{\sqrt{\pi \left[ (\mathcal{E}_s/2) \sum_{k=-\infty}^{\infty} g^2(l/(2WL) - k/(2W)) + \sigma^2 W/2 \right]}}, \end{aligned} \quad (57)$$

for  $l = -L + 1, \dots, L - 1$ .

The undistorted received signal samples,  $Y_{0,u}$  and  $Y_{0,l}$ , are jointly zero-mean Gaussian. We can further evaluate their correlation as

$$\begin{aligned} r_{u,l} &= \frac{\mathbf{E}[Y_{0,u} Y_{0,l}]}{\sqrt{\text{var}[Y_{0,u}]} \cdot \sqrt{\text{var}[Y_{0,l}]}} \\ &= \frac{\frac{\mathcal{E}_s}{2W} \sum_{k=-\infty}^{\infty} g(l/(2WL) - k/(2W)) g(u/(2WL) - k/(2W)) + \frac{\sigma^2}{2} \text{sinc}((l-u)/L)}{\sqrt{\frac{\mathcal{E}_s}{2W} \sum_{k=-\infty}^{\infty} g^2(l/(2WL) - k/(2W)) + \frac{\sigma^2}{2}} \sqrt{\frac{\mathcal{E}_s}{2W} \sum_{k=-\infty}^{\infty} g^2(u/(2WL) - k/(2W)) + \frac{\sigma^2}{2}}}. \end{aligned}$$

Consequently, the correlation between the hard-limited samples is [13]

$$\Omega_{u,l} = \mathbf{E}[W_{0,u} W_{0,l}] = \frac{2}{\pi} \arcsin r_{u,l}. \quad (58)$$

Now in this subsection we focus on the sinc pulse function,  $g(t) = \sqrt{2W} \text{sinc}(2Wt)$ . For this  $g(\cdot)$ , through (57) and (58) we have

$$b_l = \frac{2\mathcal{E}_s}{\sqrt{\pi\sigma^2}} \frac{\text{sinc}(l/L)}{\sqrt{(2\mathcal{E}_s/\sigma^2)\Xi(l,l) + 1}}, \quad (59)$$

$$r_{u,l} = \frac{(2\mathcal{E}_s/\sigma^2)\Xi(l,u) + \text{sinc}\left(\frac{l-u}{L}\right)}{\sqrt{(2\mathcal{E}_s/\sigma^2)\Xi(l,l) + 1} \sqrt{(2\mathcal{E}_s/\sigma^2)\Xi(u,u) + 1}}, \quad (60)$$

where  $\Xi(l,u) = \sum_{k=-\infty}^{\infty} \text{sinc}(l/L - k) \text{sinc}(u/L - k)$ , which can be further evaluated as  $\Xi(l,u) = \text{sinc}((l-u)/L)$ , for all  $l, u = -L + 1, \dots, L - 1$ . So

$$b_l = \sqrt{\frac{2\mathcal{E}_s}{\pi}} \sqrt{\frac{2\mathcal{E}_s/\sigma^2}{2\mathcal{E}_s/\sigma^2 + 1}} \text{sinc}(l/L), \quad \text{and } r_{u,l} = \text{sinc}\left(\frac{l-u}{L}\right). \quad (61)$$

When  $L = 1$ , i.e., Nyquist sampling, we can easily verify that  $\Delta = \frac{2}{\pi} \frac{\mathcal{E}_s}{\mathcal{E}_s + \sigma^2/2}$ , thus revisiting the result in Section IV.

From the above, we can find the following behavior of the GMI, in which we denote  $\text{SNR} = \frac{\varepsilon_a}{\sigma^2/2}$ ,  $\underline{b}_0 = [\text{sinc}(l/L)]_{l=-L+1, \dots, L-1}$ , and  $\underline{\Omega}_0 = [\arcsin \text{sinc}((l-u)/L)]_{l,u=-L+1, \dots, L-1}$ .

$$\Delta = \frac{\text{SNR}}{\text{SNR} + 1} \underline{b}_0^T \underline{\Omega}_0^{-1} \underline{b}_0, \quad \text{and} \quad \text{SNR}_e = \frac{\underline{b}_0^T \underline{\Omega}_0^{-1} \underline{b}_0 \cdot \text{SNR}}{(1 - \underline{b}_0^T \underline{\Omega}_0^{-1} \underline{b}_0) \cdot \text{SNR} + 1}. \quad (62)$$

- High-SNR regime: As  $\text{SNR} \rightarrow \infty$ ,

$$I_{\text{GMI}} = \frac{1}{2} \log \left( \frac{1}{1 - \underline{b}_0^T \underline{\Omega}_0^{-1} \underline{b}_0} \right) + o(1). \quad (63)$$

- Low-SNR regime: As  $\text{SNR} \rightarrow 0$ ,

$$I_{\text{GMI}} = \frac{\underline{b}_0^T \underline{\Omega}_0^{-1} \underline{b}_0}{2} \text{SNR} + o(\text{SNR}). \quad (64)$$

In Table IV, we present the numerical results for the asymptotic behavior of the GMI, for different values of  $L$ . From the numerical results, we see that super-Nyquist sampling yields noticeable benefit for the GMI. In the low-SNR regime, sampling at twice the Nyquist rate attains  $\lim_{\text{SNR} \rightarrow 0} I_{\text{GMI}}/\text{SNR} = 0.3587$ , which is slightly smaller than the lower bound 0.3732 which has been obtained in [7]. In the high-SNR regime, we further observe that for  $L \geq 4$  the GMI exceeds 1 bit/c.u.! Intuitively, this is due to the fact that the diversity yielded by super-Nyquist sampling is capable of exploiting the abundant information carried by the Gaussian codebook ensemble.

To further consolidate our above analysis, in Figure 4 we plot the GMI achieved for different values of  $L$ . We can clearly observe the rate gain by increasing the sampling rate. For comparison, we also plot the AWGN capacity without distortion and the capacity under binary symmetric quantization and with Nyquist sampling [5]. We notice that, as  $L$  increases, on one hand, the performance gap between the GMI and the capacity tends to diminish for SNR smaller than 0 dB; on the other hand, the GMI even outperforms the capacity at high SNR.

### B. Binary Symmetric Quantization: Pulse Function Optimization at Low SNR

We have already seen in the previous subsection that super-Nyquist sampling yields noticeable benefit. In this subsection, we illustrate that we can even realize additional benefit through optimizing the pulse function  $g(\cdot)$ .

With sampling factor  $L$ , we restrict the pulse function to take the following form

$$g(t) = \sum_{v=-L+1}^{L-1} \gamma_v \sqrt{2W} \text{sinc}(2Wt - v/L); \quad (65)$$

that is, a superposition of  $2L-1$  (time-shifted) sinc pulses. The weighting parameters  $\{\gamma_v\}_{v=-L+1}^{L-1}$  are such that the energy of  $g(t)$  is unity, *i.e.*,

$$\int_{-\infty}^{\infty} g^2(t) dt = \sum_{v=-L+1}^{L-1} \sum_{v'=-L+1}^{L-1} \gamma_v \gamma_{v'} \text{sinc}\left(\frac{v-v'}{L}\right) = 1, \quad (66)$$

which may be rewritten in matrix form as  $\underline{\gamma}^T \mathbf{\Theta} \underline{\gamma} = 1$ , where  $\mathbf{\Theta} = [\text{sinc}((l-u)/L)]_{l,u=-L+1,\dots,L-1}$ . If we let  $\gamma_0 = 1$  and  $\gamma_{v \neq 0} = 0$ , we obtain the sinc pulse function.

Through the general formulas (57) and (58), we have, after some algebraic manipulation,

$$b_l = \sqrt{\frac{2\mathcal{E}_s}{\pi}} \sqrt{\frac{2\mathcal{E}_s/\sigma^2}{2\mathcal{E}_s/\sigma^2 + 1}} \sum_{v=-L+1}^{L-1} \gamma_v \text{sinc}\left(\frac{l-v}{L}\right), \quad (67)$$

$$r_{u,l} = \frac{(2\mathcal{E}_s/\sigma^2) \sum_{a=-L+1}^{L-1} \sum_{b=-L+1}^{L-1} \gamma_a \gamma_b \text{sinc}\left(\frac{l-u-a+b}{L}\right) + \text{sinc}\left(\frac{l-u}{L}\right)}{2\mathcal{E}_s/\sigma^2 + 1}. \quad (68)$$

To illustrate the benefit of optimizing the pulse function, we focus on the low-SNR regime, where  $\text{SNR} = \frac{\mathcal{E}_s}{\sigma^2/2}$  approaches toward zero. We thus have

$$\sqrt{\frac{\pi}{2\mathcal{E}_s}} \frac{b_l}{\sqrt{\text{SNR}}} \rightarrow \sum_{v=-L+1}^{L-1} \gamma_v \text{sinc}\left(\frac{l-v}{L}\right), \quad \text{and} \quad r_{u,l} \rightarrow \text{sinc}\left(\frac{l-u}{L}\right). \quad (69)$$

Subsequently, the value of  $\Delta$  and  $\text{SNR}_e$  in Proposition 2 behaves like

$$\lim_{\text{SNR} \rightarrow 0} \frac{\text{SNR}_e}{\text{SNR}} = \lim_{\text{SNR} \rightarrow 0} \frac{\Delta}{\text{SNR}} = \underline{\gamma}^T \mathbf{\Theta} \mathbf{\Omega}_0^{-1} \mathbf{\Theta} \underline{\gamma}, \quad (70)$$

where  $\mathbf{\Theta} = [\text{sinc}((l-u)/L)]_{l,u=-L+1,\dots,L-1}$  and  $\mathbf{\Omega}_0 = [\arcsin \text{sinc}((l-u)/L)]_{l,u=-L+1,\dots,L-1}$  have been defined previously. Keeping in mind the unit-energy constraint on  $g(t)$ , the following optimization problem is immediate,

$$\max_{\underline{\gamma}} \underline{\gamma}^T \mathbf{\Theta} \mathbf{\Omega}_0^{-1} \mathbf{\Theta} \underline{\gamma}, \quad \text{s.t.} \quad \underline{\gamma}^T \mathbf{\Theta} \underline{\gamma} = 1. \quad (71)$$

By noting that  $\mathbf{\Theta}$  is a positive-definite matrix, we can introduce the transform  $\tilde{\underline{\gamma}} = \mathbf{\Theta}^{1/2} \underline{\gamma}$ , and rewrite the optimization problem as

$$\max_{\underline{\gamma}} \frac{\tilde{\underline{\gamma}}^T \mathbf{\Theta}^{1/2} \mathbf{\Omega}_0^{-1} \mathbf{\Theta}^{1/2} \tilde{\underline{\gamma}}}{\tilde{\underline{\gamma}}^T \tilde{\underline{\gamma}}}, \quad (72)$$

for which the maximum value is the largest eigenvalue of  $\mathbf{\Theta}^{1/2} \mathbf{\Omega}_0^{-1} \mathbf{\Theta}^{1/2}$ , and the optimal  $\tilde{\underline{\gamma}}$  is the unit-norm eigenvector corresponding to the largest eigenvalue.

In Table V, we present the numerical results for the low-SNR asymptotic behavior of the GMI, with the optimal choice of  $\underline{\gamma}$ , for different values of  $L$ . Compared with Table IV, we

notice that optimizing the pulse function leads to a noticeable additional improvement on the GMI. In particular, for  $L = 2$  our approach yields  $\lim_{\text{SNR} \rightarrow 0} I_{\text{GMI}}/\text{SNR} = 0.3731$ , which almost coincides with the result in [7], 0.3732.<sup>6</sup>

## VII. CONCLUSIONS

With the surging quest for energy-efficient communication solutions, transceivers with deliberately engineered distortions have attracted much attention in system design. These distortions, such as transmit-side clipping and low-precision receive-side quantization, may significantly alleviate power consumption and hardware cost. It is thus imperative for communication engineers to develop a systematic understanding of the impact of these distortions, so as to assess the resulting system performance, and to guide the design of distortion mechanisms. In this paper, we make an initial attempt at this goal, developing a general analytical framework for evaluating the achievable information rates using the measure of GMI, and illustrating the application of this framework by examining several representative transceiver distortion models. We hope that both the framework and the applications presented in this paper will be useful for deepening our understanding in this area.

Admittedly, the approach taken in this paper, namely evaluating the GMI for Gaussian codebook ensemble and nearest-neighbor decoding, is inherently suboptimal for general transceiver distortion models. Nevertheless, as illustrated throughout this paper, the general analytical framework built upon such an approach is convenient for performance evaluation and instrumental for system design. In many practically important scenarios, for example the low/moderate-SNR regime, this approach leads to near-optimal performance. Furthermore, as suggested by our analysis of super-Nyquist sampling, we can substantially alleviate the performance loss by sampling the channel output at rates higher than the Nyquist rate.

A number of interesting problems remain unsolved within the scope of this paper. These include, among others: answering whether the GMI coincides with the channel capacity for multi-bit output quantization in the low-SNR limit; identifying more effective ways of processing the samples in super-Nyquist sampled channels; characterizing the ultimate performance limit of

<sup>6</sup>Since both our result and that in [7] are analytical, we have compared their values in fine precision and found that they are indeed different.

super-Nyquist sampling. Beyond the scope of this paper, one can readily see a whole agenda of research on communication with nonlinear transceiver distortion, including timing recovery, channel estimation, equalization, transmission under multipath fading, and multiantenna/multiuser aspects.

## REFERENCES

- [1] E. Biglieri, J. Proakis, and S. Shamai (Shitz), "Fading channels: information-theoretic and communications aspects," *IEEE Trans. Inform. Theory*, Vol. 44, No. 6, pp. 2619-2692, Oct. 1998.
- [2] H. Ochiai and H. Imai, "Performance analysis of deliberately clipped OFDM signals," *IEEE Trans. Commun.*, Vol. 50, No. 1, pp. 89-101, Jan. 2002.
- [3] J. J. Bussgang, "Crosscorrelation functions of amplitude-distorted Gaussian signals," Technical Report No. 216, Research Laboratory of Electronics, Massachusetts Institute of Technology, Cambridge, MA, Mar. 1952.
- [4] A. J. Viterbi and J. K. Omura, *Principles of Digital Communication and Coding*, McGraw-Hill, 1979.
- [5] J. Singh, O. Dabeer, and U. Madhow, "On the limits of communication with low-precision analog-to-digital conversion at the receiver," *IEEE Trans. Commun.*, Vol. 57, No. 12, pp. 3629-3639, Dec. 2009.
- [6] A. Mezghani and J. A. Nossek, "Analysis of Rayleigh-fading channels with 1-bit quantized output," in *Proc. IEEE Int. Symp. Inform. Theory (ISIT)*, pp. 260-264, Toronto, Canada, Jul. 2008.
- [7] T. Koch and A. Lapidoth, "Increased capacity per unit-cost by oversampling," [Online] <http://arxiv.org/abs/1008.5393v1>
- [8] T. Koch and A. Lapidoth, "Does output quantization really cause a loss of 2dB?" [Online] <http://arxiv.org/abs/1101.0970v1>
- [9] A. Ganti, A. Lapidoth, and I. E. Telatar, "Mismatched decoding revisited: General alphabets, channels with memory, and the wide-band limit," *IEEE Trans. Inform. Theory*, Vol. 46, No. 7, pp. 2315-2328, Nov. 2000.
- [10] A. Lapidoth and S. Shamai (Shitz), "Fading channels: How perfect need 'perfect side information' be?" *IEEE Trans. Inform. Theory*, Vol. 48, No. 5, pp. 1118-1134, May 2002.
- [11] J. Salz and E. Zehavi, "Decoding under integer metrics constraints," *IEEE Trans. Commun.*, Vol. 43, No. 2-4, pp. 307-317, Feb.-Apr. 1995.
- [12] S. Verdú, "On channel capacity per unit cost," *IEEE Trans. Inform. Theory*, Vol. 36, No. 5, pp. 1019-1030, Sep. 1990.
- [13] J. H. Van Vleck and D. Middleton, "The spectrum of clipping noise," *Proc. IEEE*, Vol. 54, No. 1, pp. 2-19, Jan. 1966.

## SUPPLEMENTARY MATERIAL

### A. Derivation of the GMI in Proposition 1

We proceed starting from (6) as follows. For any  $m \neq 1$ ,

$$\begin{aligned}
 \mathbf{E} \left\{ e^{n\theta D(m)} \middle| \mathbf{W}_k, k = 1, \dots, n \right\} &= \mathbf{E} \left\{ e^{\theta \sum_{k=1}^n [\mathbf{W}_k - a\mathbf{X}_k(m)]^2} \middle| \mathbf{W}_k, k = 1, \dots, n \right\} \\
 &= \prod_{k=1}^n \mathbf{E} \left\{ e^{\theta [\mathbf{W}_k - a\mathbf{X}_k(m)]^2} \middle| \mathbf{W}_k \right\} = \prod_{k=1}^n \frac{1}{\sqrt{1 - 2\theta a^2 \mathcal{E}_s}} \exp \left( \frac{\theta \mathbf{W}_k^2}{1 - 2\theta a^2 \mathcal{E}_s} \right) \\
 &= (1 - 2\theta a^2 \mathcal{E}_s)^{-n/2} \exp \left( \sum_{k=1}^n \frac{\theta \mathbf{W}_k^2}{1 - 2\theta a^2 \mathcal{E}_s} \right), \tag{73}
 \end{aligned}$$

by noting that conditioned upon  $\mathbf{W}$ ,  $(\mathbf{W} - a\mathbf{X})^2$  is a noncentral chi-square random variable.

This leads to

$$\Lambda_n(n\theta) = \log \mathbf{E} \left\{ e^{n\theta D(m)} \middle| \mathbf{W}_k, k = 1, \dots, n \right\} = \frac{\theta}{1 - 2\theta a^2 \mathcal{E}_s} \sum_{k=1}^n \mathbf{W}_k^2 - \frac{n}{2} \log(1 - 2\theta a^2 \mathcal{E}_s). \tag{74}$$

Consequently, from the law of large numbers,

$$\Lambda(\theta) = \lim_{n \rightarrow \infty} \frac{1}{n} \Lambda_n(n\theta) = \frac{\theta \mathbf{E}[f(\mathbf{X}, \mathbf{Z})]^2}{1 - 2\theta a^2 \mathcal{E}_s} - \frac{1}{2} \log(1 - 2\theta a^2 \mathcal{E}_s) \quad \text{a.s.} \tag{75}$$

where  $\mathbf{X} \sim \mathcal{N}(0, \mathcal{E}_s)$  and  $\mathbf{Z} \sim \mathcal{N}(0, \sigma^2)$ . So we can evaluate the GMI through

$$I_{\text{GMI}} = \sup_{a \in \mathbb{R}, \theta < 0} \left( \theta \mathbf{E} \left\{ [f(\mathbf{X}, \mathbf{Z}) - a\mathbf{X}]^2 \right\} - \frac{\theta \mathbf{E}[f(\mathbf{X}, \mathbf{Z})]^2}{1 - 2\theta a^2 \mathcal{E}_s} + \frac{1}{2} \log(1 - 2\theta a^2 \mathcal{E}_s) \right). \tag{76}$$

Note that in the problem formulation we include the optimization of  $I_{\text{GMI}}$  over  $a \in \mathbb{R}$ .

To solve the optimization problem, we define

$$\begin{aligned}
 J(a, \theta) &= \theta \mathbf{E} \left\{ [f(\mathbf{X}, \mathbf{Z}) - a\mathbf{X}]^2 \right\} - \frac{\theta \mathbf{E}[f(\mathbf{X}, \mathbf{Z})]^2}{1 - 2\theta a^2 \mathcal{E}_s} + \frac{1}{2} \log(1 - 2\theta a^2 \mathcal{E}_s) \\
 &= \theta \left\{ \mathbf{E}[f(\mathbf{X}, \mathbf{Z})]^2 + a^2 \mathcal{E}_s - 2a \mathbf{E}[f(\mathbf{X}, \mathbf{Z})\mathbf{X}] \right\} - \frac{\theta \mathbf{E}[f(\mathbf{X}, \mathbf{Z})]^2}{1 - 2\theta a^2 \mathcal{E}_s} + \frac{1}{2} \log(1 - 2\theta a^2 \mathcal{E}_s) \\
 &= \theta a^2 \mathcal{E}_s + \frac{1}{2} \log(1 - 2\theta a^2 \mathcal{E}_s) - \frac{2\theta^2 a^2 \mathcal{E}_s}{1 - 2\theta a^2 \mathcal{E}_s} \mathbf{E}[f(\mathbf{X}, \mathbf{Z})]^2 - 2\theta a \mathbf{E}[f(\mathbf{X}, \mathbf{Z})\mathbf{X}]. \tag{77}
 \end{aligned}$$

By introducing the new variable  $\gamma = -2\theta a^2 \mathcal{E}_s > 0$ , we rewrite  $J(a, \theta)$  as

$$J(\gamma, \theta) = \frac{1}{2} \log(1 + \gamma) - \frac{\gamma}{2} + \frac{\gamma \theta}{1 + \gamma} \mathbf{E}[f(\mathbf{X}, \mathbf{Z})]^2 + \sqrt{\frac{-2\gamma \theta}{\mathcal{E}_s}} \mathbf{E}[|f(\mathbf{X}, \mathbf{Z})\mathbf{X}|]. \tag{78}$$

Letting the partial derivative  $\partial J / \partial \theta$  be zero, we find that the optimal value of  $\theta < 0$  should be

$$\sqrt{-\theta_{\text{opt}}} = \frac{(1 + \gamma) \mathbf{E}[|f(\mathbf{X}, \mathbf{Z})\mathbf{X}|]}{\mathbf{E}[f(\mathbf{X}, \mathbf{Z})]^2 \sqrt{2\mathcal{E}_s \gamma}}. \tag{79}$$

Substituting  $\theta_{\text{opt}}$  into  $J(\gamma, \theta)$  followed by some algebraic manipulation, we obtain

$$J(\gamma, \theta_{\text{opt}}) = \frac{1}{2} \log(1 + \gamma) - \frac{\gamma}{2} + \frac{(1 + \gamma) \{\mathbf{E}[f(X, Z)X]\}^2}{2\mathcal{E}_s \mathbf{E}[f(X, Z)]^2}. \quad (80)$$

Let us define

$$\Delta = \frac{\{\mathbf{E}[f(X, Z)X]\}^2}{\mathcal{E}_s \mathbf{E}[f(X, Z)]^2}, \quad (81)$$

and maximize  $J(\gamma, \theta_{\text{opt}}) = \frac{1}{2} \log(1 + \gamma) - \frac{\gamma}{2} + (1 + \gamma) \frac{\Delta}{2}$  over  $\gamma > 0$ . From Cauchy-Schwartz inequality, we see that  $\Delta$  is upper bounded by one. It is then straightforward to show that the optimal value of  $\gamma$  is  $\gamma_{\text{opt}} = \Delta/(1 - \Delta)$ , and hence  $J(\gamma_{\text{opt}}, \theta_{\text{opt}}) = -\frac{1}{2} \log(1 - \Delta)$ .

Therefore, the maximum value  $J(\gamma_{\text{opt}}, \theta_{\text{opt}})$ , *i.e.*, the GMI, is

$$I_{\text{GMI}} = \frac{1}{2} \log \left( 1 + \frac{\Delta}{1 - \Delta} \right), \quad (82)$$

and the optimal choice of the decoding scaling parameter  $a$  is  $a_{\text{opt}} = \mathbf{E}[f(X, Z)X] / \mathcal{E}_s$ .

### B. Derivation of the GMI for Antipodal Codebook Ensemble

We follow the same line of analysis as that for the Gaussian codebook ensemble. For  $m = 1$ ,

$$\begin{aligned} \lim_{n \rightarrow \infty} D(1) &= \mathbf{E} \{ [W - aX]^2 \} \\ &= \mathbf{E}[W^2] + a^2 \mathcal{E}_s - 2a \mathbf{E}[WX] \quad \text{a.s.} \end{aligned} \quad (83)$$

where  $W = f(X, Z)$  denotes the distorted channel output. On the other hand, for any  $m \neq 1$ , we find that

$$\frac{1}{n} \Lambda_n(n\theta) = \frac{\theta}{n} \sum_{k=1}^n W_k^2 + \theta a^2 \mathcal{E}_s + \frac{1}{n} \sum_{k=1}^n \log \cosh(2\theta a \sqrt{\mathcal{E}_s} W_k), \quad (84)$$

$$\text{and } \Lambda(\theta) = \lim_{n \rightarrow \infty} \frac{1}{n} \Lambda_n(n\theta) = \theta \mathbf{E}[W^2] + \theta a^2 \mathcal{E}_s + \mathbf{E} \log \cosh(2\theta a \sqrt{\mathcal{E}_s} W), \quad \text{a.s.} \quad (85)$$

Consequently, we can evaluate the GMI by solving

$$I_{\text{GMI}} = \sup_{\theta < 0, a \in \mathbb{R}} \left( -2\theta a \mathbf{E}[Xf(X, Z)] - \mathbf{E} \log \cosh(2\theta a \sqrt{\mathcal{E}_s} f(X, Z)) \right). \quad (86)$$

By letting  $-2\theta a$  be a single variable  $t$ , we obtain the problem formulation as given by (9), and by using the first derivative condition for optimality, we obtain the equation for the optimal value of  $t$  as given by (10).

### C. General Framework for Complex-Valued Nyquist-Sampled Channels

We can extend the general GMI formula (7) for real-valued channels to complex-valued channels. Let the noise  $\mathbf{Z}$  be a sequence of i.i.d. circularly symmetric complex Gaussian random variables (*i.e.*,  $\mathbf{Z} \sim \mathcal{CN}(0, \sigma^2)$ ). The memoryless nonlinearity mapping  $f(\cdot)$  transforms  $(x, z)$  into a complex number  $f(x, z)$ . Hence the observation is  $W_k = f(X_k, Z_k)$ , for  $k = 1, 2, \dots, n$ .

For transmission, we restrict the codebook to be an i.i.d.  $\mathcal{CN}(0, \mathcal{E}_s)$  ensemble. The decoder follows a nearest-neighbor rule, which computes for all possible messages, the distance metric,

$$D(m) = \frac{1}{n} \sum_{k=1}^n |w_k - ax_k(m)|^2, \quad m \in \mathcal{M}, \quad (87)$$

and decides the received message as  $\hat{m} = \arg \min_{m \in \mathcal{M}} D(m)$ .

Analogously to the development for the real-valued channel model in Section II, we arrive at

$$I_{\text{GMI}} = \sup_{a \in \mathbb{C}, \theta < 0} \left( \theta \mathbf{E} \{ |f(X, Z) - aX|^2 \} - \frac{\theta \mathbf{E} |f(X, Z)|^2}{1 - \theta |a|^2 \mathcal{E}_s} + \log(1 - \theta |a|^2 \mathcal{E}_s) \right). \quad (88)$$

Note that in the problem formulation we include the optimization of  $I_{\text{GMI}}$  over  $a \in \mathbb{C}$ .

Define the expression in the right-hand side of (88) as  $J(a, \theta)$ , which can be rewritten as

$$J(a, \theta) = \theta |a|^2 \mathcal{E}_s + \log(1 - \theta |a|^2 \mathcal{E}_s) - \frac{\theta^2 |a|^2 \mathcal{E}_s \mathbf{E} |f(X, Z)|^2}{1 - \theta |a|^2 \mathcal{E}_s} - 2\theta |a| \mathcal{R} \mathbf{E} \{ e^{j\phi} \bar{f}(X, Z) X \}, \quad (89)$$

where  $\phi$  is the phase of  $a$ , and  $\mathcal{R}$  denotes the real part of a complex number. By introducing the new variable  $\gamma = -\theta |a|^2 \mathcal{E}_s > 0$ , we further rewrite  $J(a, \theta)$  as

$$J(\gamma, \phi, \theta) = \log(1 + \gamma) - \gamma + \frac{\gamma \theta}{1 + \gamma} \mathbf{E} |f(X, Z)|^2 + 2\sqrt{\frac{-\gamma \theta}{\mathcal{E}_s}} \mathcal{R} \mathbf{E} \{ e^{j\phi} \bar{f}(X, Z) X \}. \quad (90)$$

Letting the partial derivative  $\partial J / \partial \theta$  be zero, we find that the optimal value of  $\theta < 0$  should be

$$\sqrt{-\theta_{\text{opt}}} = \frac{(1 + \gamma) \mathcal{R} \mathbf{E} \{ e^{j\phi} \bar{f}(X, Z) X \}}{\mathbf{E} |f(X, Z)|^2 \sqrt{\mathcal{E}_s \gamma}}. \quad (91)$$

Substituting  $\theta_{\text{opt}}$  into  $J(\gamma, \theta)$  followed by some algebraic manipulation, we obtain

$$J(\gamma, \phi, \theta_{\text{opt}}) = \log(1 + \gamma) - \gamma + \frac{(1 + \gamma) [\mathcal{R} \mathbf{E} \{ e^{j\phi} \bar{f}(X, Z) X \}]^2}{\mathcal{E}_s \mathbf{E} |f(X, Z)|^2}. \quad (92)$$

Let us define

$$\Delta(\phi) = \frac{[\mathcal{R} \mathbf{E} \{ e^{j\phi} \bar{f}(X, Z) X \}]^2}{\mathcal{E}_s \mathbf{E} |f(X, Z)|^2}, \quad (93)$$



and maximize  $J(\gamma, \phi, \theta_{\text{opt}}) = \log(1 + \gamma) - \gamma + (1 + \gamma)\Delta(\phi)$  over  $\gamma > 0$ . It is straightforward to show that the optimal value of  $\gamma$  is  $\gamma_{\text{opt}} = \frac{\Delta(\phi)}{1 - \Delta(\phi)}$ , and hence  $J(\gamma_{\text{opt}}, \phi, \theta_{\text{opt}}) = -\log(1 - \Delta(\phi))$ .

It is clear that  $J(\gamma_{\text{opt}}, \phi, \theta_{\text{opt}})$  is maximized by choosing  $\phi = \phi_{\text{opt}} = -\arctan \mathbf{E} \{ \bar{f}(\mathbf{X}, \mathbf{Z})\mathbf{X} \}$ , which maximizes  $\Delta(\phi)$ . Denote  $\Delta(\phi_{\text{opt}})$  by  $\Delta_{\text{opt}}$ , which is

$$\Delta_{\text{opt}} = \frac{|\mathbf{E} \{ \bar{f}(\mathbf{X}, \mathbf{Z})\mathbf{X} \}|^2}{\mathcal{E}_s \mathbf{E} |f(\mathbf{X}, \mathbf{Z})|^2}. \quad (94)$$

Therefore, the maximum value  $J(\gamma_{\text{opt}}, \phi_{\text{opt}}, \theta_{\text{opt}})$ , i.e., the GMI, is

$$I_{\text{GMI}} = J(\gamma_{\text{opt}}, \phi_{\text{opt}}, \theta_{\text{opt}}) = \log \left( 1 + \frac{\Delta_{\text{opt}}}{1 - \Delta_{\text{opt}}} \right) = \log(1 + \text{SNR}_e), \quad (95)$$

and the optimal choice of the decoding scaling parameter  $a$  is  $a_{\text{opt}} = \mathbf{E} \{ f(\mathbf{X}, \mathbf{Z})\bar{\mathbf{X}} \} / \mathcal{E}_s$ .

#### D. Derivation of Eqn. (25) and (26)

$$\begin{aligned} \mathbf{E}[f(\mathbf{X} + \mathbf{Z})]^2 &= 2 \sum_{i=1}^M \iint_{\alpha_{i-1} \leq x+z < \alpha_i} r_i^2 p_{\mathbf{X}}(x) p_{\mathbf{Z}}(z) dx dz \\ &= 2 \sum_{i=1}^M r_i^2 \int_{\alpha_{i-1}}^{\alpha_i} \frac{\exp\left(-\frac{y^2}{2(\mathcal{E}_s + \sigma^2)}\right)}{\sqrt{2\pi(\mathcal{E}_s + \sigma^2)}} dy = 2 \sum_{i=1}^M r_i^2 \left[ Q\left(\frac{\alpha_{i-1}}{\sqrt{\mathcal{E}_s + \sigma^2}}\right) - Q\left(\frac{\alpha_i}{\sqrt{\mathcal{E}_s + \sigma^2}}\right) \right], \\ \mathbf{E}[f(\mathbf{X} + \mathbf{Z})\mathbf{X}] &= 2 \sum_{i=1}^M \iint_{\alpha_{i-1} \leq x+z < \alpha_i} r_i x p_{\mathbf{X}}(x) p_{\mathbf{Z}}(z) dx dz \\ &= 2 \sum_{i=1}^M r_i \int_{-\infty}^{\infty} p_{\mathbf{Z}}(z) \left( \int_{\alpha_{i-1}-z}^{\alpha_i-z} x p_{\mathbf{X}}(x) dx \right) dz \\ &= 2 \sum_{i=1}^M r_i \left[ \int_{-\infty}^{\infty} p_{\mathbf{Z}}(z) F(\alpha_{i-1} - z) dz - \int_{-\infty}^{\infty} p_{\mathbf{Z}}(z) F(\alpha_i - z) dz \right] \\ &= \mathcal{E}_s \sqrt{\frac{2}{\pi(\mathcal{E}_s + \sigma^2)}} \sum_{i=1}^M r_i \left[ e^{-\frac{\alpha_{i-1}^2}{2(\mathcal{E}_s + \sigma^2)}} - e^{-\frac{\alpha_i^2}{2(\mathcal{E}_s + \sigma^2)}} \right]. \end{aligned}$$

#### E. Nearest-Neighbor Decoding for Antipodal Input and Symmetric Output Quantizers

For a given  $2M$ -level symmetric quantizer, and for antipodal inputs, we can evaluate the GMI following the result in Section II. Denote the probability  $\Pr[W = r_i | \mathbf{X} = \sqrt{\mathcal{E}_s}]$  by  $p_i^{(+)}$  and  $\Pr[W = -r_i | \mathbf{X} = \sqrt{\mathcal{E}_s}]$  by  $p_i^{(-)}$ ; by symmetry, we have  $\Pr[W = r_i | \mathbf{X} = -\sqrt{\mathcal{E}_s}] = p_i^{(-)}$  and

$\Pr[W = -r_i | X = -\sqrt{\mathcal{E}_s}] = p_i^{(+)}$ , and  $\Pr[W = r_i] = \Pr[W = -r_i] = (p_i^{(+)} + p_i^{(-)})/2$ . The GMI thus is

$$I_{\text{GMI}} = \sup_{t \in \mathbb{R}} \left( t \sqrt{\mathcal{E}_s} \sum_{i=1}^M (p_i^{(+)} - p_i^{(-)}) r_i - \sum_{i=1}^M (p_i^{(+)} + p_i^{(-)}) \log \cosh(t \sqrt{\mathcal{E}_s} r_i) \right). \quad (96)$$

Maximizing GMI with respect to the reconstruction points  $\underline{r}$ , we have that the optimal  $\underline{r}$  satisfies

$$r_i = \frac{1}{t \sqrt{\mathcal{E}_s}} \text{artanh} \left( \frac{p_i^{(+)} - p_i^{(-)}}{p_i^{(+)} + p_i^{(-)}} \right) = \frac{1}{2t \sqrt{\mathcal{E}_s}} \log \frac{p_i^{(+)}}{p_i^{(-)}}, \quad i = 1, \dots, M, \quad (97)$$

and that the GMI further reduces into

$$\begin{aligned} I_{\text{GMI}} &= \sum_{i=1}^M \left[ \frac{p_i^{(+)} - p_i^{(-)}}{2} \log \frac{p_i^{(+)}}{p_i^{(-)}} + (p_i^{(+)} + p_i^{(-)}) \log 2 - (p_i^{(+)} + p_i^{(-)}) \log \left( \sqrt{\frac{p_i^{(+)}}{p_i^{(-)}}} + \sqrt{\frac{p_i^{(-)}}{p_i^{(+)}}} \right) \right] \\ &= \log 2 - \sum_{i=1}^M \left[ (p_i^{(+)} + p_i^{(-)}) \log(p_i^{(+)} + p_i^{(-)}) - p_i^{(+)} \log p_i^{(+)} - p_i^{(-)} \log p_i^{(-)} \right] = I(\mathbf{X}; \mathbf{W}). \end{aligned} \quad (98)$$

That is, the GMI coincides with the channel input-output mutual information, which is achievable by maximum-likelihood decoding. This seemingly surprising result is in fact reasonable, because there is indeed a nearest-neighbor decoding realization of the maximum-likelihood decoding rule, when the channel input is antipodal and the output quantization is symmetric. Choosing the reconstruction points as  $r_i = \log[p_i^{(+)} / p_i^{(-)}]$ ,  $i = 1, \dots, M$ , and denoting  $w_k$  by  $r_{w_k} \cdot \text{sgn}(w_k)$ , we can write the nearest-neighbor decoding metric as

$$\begin{aligned} D(m) &= \frac{1}{n} \sum_{k=1}^n \left[ \log \frac{p_{r_{w_k}}^{(+)}}{p_{r_{w_k}}^{(-)}} \text{sgn}(w_k) - x_k(m) \right]^2 \\ &= \frac{1}{n} \sum_{k=1}^n \left[ \log \frac{p_{r_{w_k}}^{(+)}}{p_{r_{w_k}}^{(-)}} \right]^2 + \mathcal{E}_s - \frac{2}{n} \sum_{k=1}^n \log \frac{p_{r_{w_k}}^{(+)}}{p_{r_{w_k}}^{(-)}} \text{sgn}(w_k) x_k(m). \end{aligned} \quad (99)$$

The first two terms in (99) are independent of the codeword, and thus it suffices to examine

$$D_1(m) = \frac{1}{n} \sum_{k=1}^n \log \frac{p_{r_{w_k}}^{(+)}}{p_{r_{w_k}}^{(-)}} \text{sgn}(w_k) x_k(m), \quad (100)$$

which can be further equivalently deduced into

$$\begin{aligned} D_2(m) &= \frac{1}{2n \sqrt{\mathcal{E}_s}} \sum_{k=1}^n \left[ \log \frac{p_{r_{w_k}}^{(+)}}{p_{r_{w_k}}^{(-)}} \text{sgn}(w_k x_k(m)) + \log(p_{r_{w_k}}^{(+)} p_{r_{w_k}}^{(-)}) \right] \\ &= \frac{1}{n \sqrt{\mathcal{E}_s}} \sum_{k=1}^n \log \Pr[w_k | x_k(m)], \end{aligned} \quad (101)$$

identical to the metric in maximum-likelihood decoding.

### F. Super-Nyquist Output Sampling with Antipodal Inputs

We examine the scenario where the input is antipodal, and where the decoder follows the linearly weighted nearest-neighbor decoding rule:

$$D(m) = \frac{1}{n} \sum_{k=1}^n \left[ \sum_{l=0}^{L-1} \beta_l w_{k,l} - x_k(m) \right]^2, \quad m \in \mathcal{M}. \quad (102)$$

Following the same line of analysis as that for the Gaussian codebook ensemble, we have, for  $m = 1$ ,

$$\lim_{n \rightarrow \infty} D(1) = \mathbf{E} \left[ \sum_{l=0}^{L-1} \beta_l W_{0,l} - X_0 \right]^2 \quad \text{a.s.} \quad (103)$$

and for any  $m \neq 1$ ,

$$\begin{aligned} \Lambda(\theta) &= \lim_{n \rightarrow \infty} \frac{1}{n} \Lambda_n(n\theta) \\ &= \theta \mathbf{E} \left[ \left( \sum_{l=0}^{L-1} \beta_l W_{0,l} \right)^2 \right] + \theta \mathcal{E}_s + \mathbf{E} \left[ \log \cosh(2\theta \sqrt{\mathcal{E}_s} \sum_{l=0}^{L-1} \beta_l W_{0,l}) \right] \quad \text{a.s.} \end{aligned} \quad (104)$$

where  $\{W_{0,l}\}_{l=0}^{L-1}$  are induced by an infinite sequence of inputs,  $\{X_k\}_{k=-\infty}^{\infty}$ . Through some manipulations, we thus obtain the resulting GMI as

$$I_{\text{GMI}} = \sup_{\underline{\beta}} \left\{ \mathbf{E} \left[ X_0 \sum_{l=0}^{L-1} \beta_l W_{0,l} \right] - \mathbf{E} \left[ \log \cosh(\sqrt{\mathcal{E}_s} \sum_{l=0}^{L-1} \beta_l W_{0,l}) \right] \right\}. \quad (105)$$

Consequently, the optimal choice of the weighting coefficients,  $\underline{\beta}$ , obeys

$$\mathbf{E} \left[ W_{0,l} \cdot \tanh \left( \sqrt{\mathcal{E}_s} \sum_{j=0}^{L-1} \beta_j W_{0,j} \right) \right] = \mathbf{E} \left[ \frac{X_0 W_{0,l}}{\sqrt{\mathcal{E}_s}} \right], \quad l = 0, 1, \dots, L-1, \quad (106)$$

which constitute an array of transcendental equations.

We further focus on the special case of binary symmetric quantizer  $w = \text{sgn}(x+z)$  and  $L = 2$ . From the symmetry in the setup, we see that  $\beta_0 = \beta_1 = \beta$ , and we only need to solve a single equation:

$$\mathbf{E}[W_{0,0} \cdot \tanh(\sqrt{\mathcal{E}_s} \beta (W_{0,0} + W_{0,1}))] = \frac{1}{\sqrt{\mathcal{E}_s}} \mathbf{E}[X_0 W_{0,0}]. \quad (107)$$

For convenience, we denote  $\Pr[(W_{0,0}, W_{0,1}) = (1, 1)] = \Pr[(W_{0,0}, W_{0,1}) = (-1, -1)] = \eta$ ,  $\Pr[(W_{0,0}, W_{0,1}) = (1, -1)] = \Pr[(W_{0,0}, W_{0,1}) = (-1, 1)] = 1/2 - \eta$ , and  $\Pr[W_{0,0} = 1 | X_0 = \sqrt{\mathcal{E}_s}] = \kappa$ . So (107) becomes

$$\tanh(2\sqrt{\mathcal{E}_s} \beta) = \frac{2\kappa - 1}{2\eta}, \quad \text{i.e.,} \quad \beta = \frac{1}{4\sqrt{\mathcal{E}_s}} \log \frac{2(\eta + \kappa) - 1}{2(\eta - \kappa) + 1}. \quad (108)$$

### G. Derivation of the GMI in Proposition 2

Denoting the expression in the right-hand side of (53) by  $J(\underline{\beta}, \theta)$ , and enforcing its partial derivatives with respect to  $\{\beta_l\}_{l=-L+1}^{L-1}$  to vanish, we have

$$\begin{aligned} \frac{\partial J}{\partial \beta_l} &= 2\theta \mathbf{E} \left[ \left( \sum_{u=-L+1}^{L-1} \beta_u W_{0,u} - X_0 \right) W_{0,l} \right] - \frac{2\theta}{1 - 2\theta \mathcal{E}_s} \mathbf{E} \left[ \left( \sum_{u=-L+1}^{L-1} \beta_u W_{0,u} \right) W_{0,l} \right] = 0 \\ \Rightarrow \sum_{u=-L+1}^{L-1} \beta_u \mathbf{E}[W_{0,u} W_{0,l}] &= \left( 1 - \frac{1}{2\theta \mathcal{E}_s} \right) \mathbf{E}[X_0 W_{0,l}], \end{aligned} \quad (109)$$

for  $l = -L+1, \dots, L-1$ . Summarizing these  $2L-1$  equations, we can write them collectively as

$$\mathbf{\Omega} \underline{\beta} = \left( 1 - \frac{1}{2\theta \mathcal{E}_s} \right) \underline{b}, \quad (110)$$

where  $\mathbf{\Omega}$  is a  $(2L-1) \times (2L-1)$  matrix with its  $(u, l)$ -element being  $\mathbf{E}[W_{0,u} W_{0,l}]$ , and  $\underline{b}$  is a  $(2L-1)$ -dimensional vector with its  $l$ -element being  $\mathbf{E}[X_0 W_{0,l}]$ . Hence we have

$$\underline{\beta} = \left( 1 - \frac{1}{2\theta \mathcal{E}_s} \right) \mathbf{\Omega}^{-1} \underline{b}. \quad (111)$$

Substituting (111) into  $J(\underline{\beta}, \theta)$ , we get

$$\begin{aligned} J(\underline{\beta}, \theta) &= \frac{2\theta^2 \mathcal{E}_s}{2\theta \mathcal{E}_s - 1} \sum_{l=-L+1}^{L-1} \sum_{u=-L+1}^{L-1} \beta_l \beta_u \Omega_{u,l} + \theta \mathcal{E}_s - 2\theta \sum_{l=-L+1}^{L-1} \beta_l b_l + \frac{1}{2} \log(1 - 2\theta \mathcal{E}_s) \\ &= \theta \mathcal{E}_s + \left( \frac{1}{2\mathcal{E}_s} - \theta \right) \underline{b}^T \mathbf{\Omega}^{-1} \underline{b} + \frac{1}{2} \log(1 - 2\theta \mathcal{E}_s). \end{aligned} \quad (112)$$

From (112), we maximize  $J(\underline{\beta}, \theta)$  by letting

$$1 - 2\theta \mathcal{E}_s = \frac{\mathcal{E}_s}{\mathcal{E}_s - \underline{b}^T \mathbf{\Omega}^{-1} \underline{b}}, \quad (113)$$

and the maximum value of  $J(\underline{\beta}, \theta)$ , i.e., the GMI, is

$$I_{\text{GMI}} = \frac{1}{2} \log \left( 1 + \frac{\underline{b}^T \mathbf{\Omega}^{-1} \underline{b} / \mathcal{E}_s}{1 - \underline{b}^T \mathbf{\Omega}^{-1} \underline{b} / \mathcal{E}_s} \right). \quad (114)$$

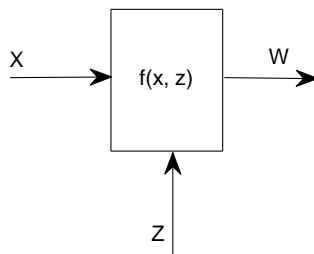


Fig. 1. Illustration of the general channel model with distortion.

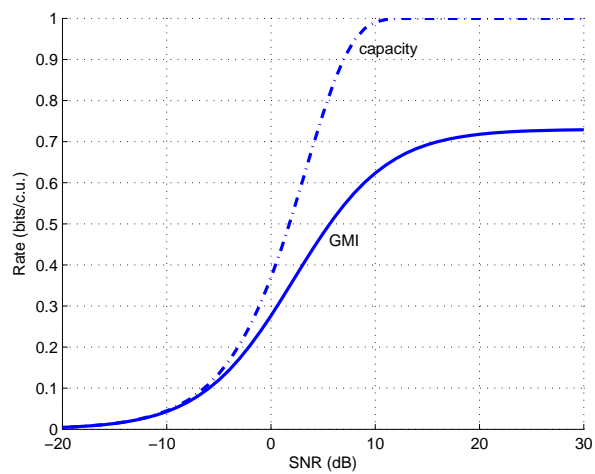


Fig. 2. The GMI and the channel capacity of the real Gaussian channel with binary symmetric output quantization.

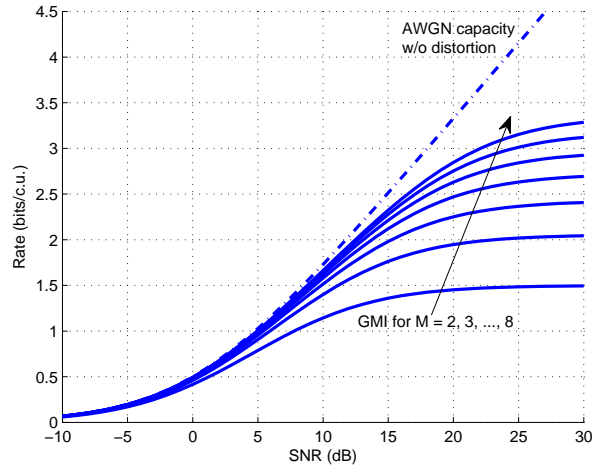


Fig. 3. The GMI achieved by optimal  $2M$ -level quantizers, for  $M = 2, 3, \dots, 8$ .

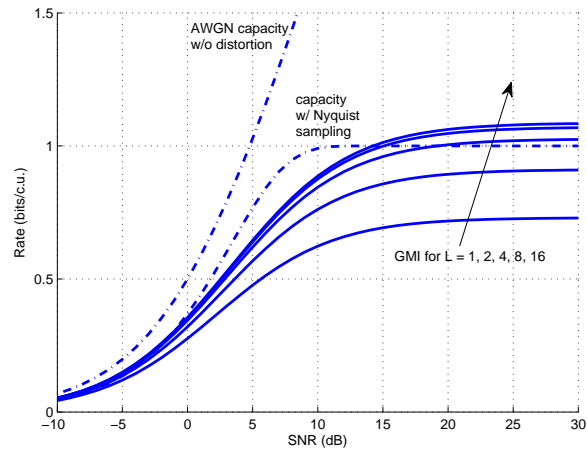


Fig. 4. The GMI achieved by super-Nyquist sampling with binary symmetric quantization and sinc pulse function, for  $L = 1, 2, 4, 8, 16$ .

$M$	2	3	4	5	6	7	8
$\max_{\alpha} K_{\underline{t}}$	2.7725	2.9569	3.0291	3.0651	3.0858	3.0989	3.1077
optimal $\alpha$	0.481	0.253	0.159	0.111	0.082	0.064	0.051

TABLE I

TABLE OF PERFORMANCE FOR OPTIMIZED UNIFORM  $2M$ -LEVEL SYMMETRIC OUTPUT QUANTIZATION.

$M$	2	3	4	5	6	7	8
$K_{\underline{t}}$	2.7488	2.9267	3.0011	3.0404	3.0642	3.0798	3.0908

TABLE II

TABLE OF PERFORMANCE FOR  $t$ -UNIFORM  $2M$ -LEVEL SYMMETRIC OUTPUT QUANTIZATION.

$M$	2	3	4	5	6	7	8
$\max_t K_{\underline{t}}$	2.7725	2.9595	3.0330	3.0695	3.0902	3.1032	3.1117
optimal $t_1$	0.618	0.805	0.880	0.922	0.943	0.958	0.967

TABLE III

TABLE OF PERFORMANCE FOR OPTIMAL  $2M$ -LEVEL SYMMETRIC OUTPUT QUANTIZATION.

$L$	1	2	4	8	16	32	$\infty$
$\underline{b}_0^T \Omega_0^{-1} \underline{b}_0$	$2/\pi$	0.7173	0.7591	0.7734	0.7783	0.7801	0.7815
$\lim_{\text{SNR} \rightarrow \infty} I_{\text{GMI}}$ (bits/c.u.)	0.7302	0.9114	1.0268	1.0710	1.0867	1.0926	1.0970
$\lim_{\text{SNR} \rightarrow 0} I_{\text{GMI}}/\text{SNR}$	0.3183	0.3587	0.3796	0.3867	0.3892	0.3901	0.3907

TABLE IV

TABLE OF PERFORMANCE FOR SUPER-NYQUIST OUTPUT SAMPLING WITH BINARY SYMMETRIC QUANTIZATION AND SINC PULSE FUNCTION.

$L$	2	4	8	16	32	$\infty$
$\lim_{\text{SNR} \rightarrow 0} I_{\text{GMI}}/\text{SNR}$	0.3731	0.3923	0.3971	0.3984	0.3987	0.3988

TABLE V

TABLE OF PERFORMANCE FOR SUPER-NYQUIST OUTPUT SAMPLING WITH BINARY SYMMETRIC QUANTIZATION AND OPTIMIZED PULSE FUNCTION.